First applications of sound-based control on a mobile robot equipped with two microphones

Aly Magassouba, Nancy Bertin, François Chaumette

First applications of sound-based control on a mobile robot equipped with two microphones

Aly Magassouba[1], Nancy Bertin[2] and François Chaumette[3]

***Abstract*— This paper validates experimentally a novel approach to robot audition, sound-based control, which consists in introducing auditory features directly as inputs of a closed-loop control scheme, that is, without any explicit localization process. The applications we present rely on the implicit bearings of the sound sources computed from the time difference of arrival (TDOA) between two microphones. By linking the motion of the robot to the aural perception of the environment, this approach has the benefit of being more robust to reverberation and noise. Therefore neither complex tracking method such as Kalman filtering nor TDOA enhancement with denoising or dereverberation methods are needed to track the correct TDOA measurements. The experiments conducted on a mobile robot instrumented with a pair of microphones show the validity of our approach. In a reverberating and noisy room, this approach is able to orient the robot to a mobile sound source in real time. A positioning task with respect to two sound sources is also performed while the robot perception is disturbed by altered and spurious TDOA measurements.**

I. INTRODUCTION

As a part of the five natural senses, aural perception provides rich information that naturally complements the other senses. Exploiting such information that uses cheap sensors, with omnidirectional properties and less occlusions, brings an increase in value in the perception of the environment. It is then natural that the interest in robot audition has raised accordingly in the robotic community. In this topic that encompasses fields like speech processing or auditory scene analysis, sound source localization (SSL) remains a challenging task that consists in computing the location of sound sources with respect to (w.r.t) the robot frame on which are embedded the sound sensors.

The current work on SSL has led to the development of several branches such as binaural hearing [17][5][12] or microphone array-based localization [10][16]. These two approaches consider an embedded auditory system in an environment of static or moving sound sources. They imply a first step of localization, followed by a tracking of the sound source(s) in the scene. The auditory system is thus completely independent from the control part. Consequently since the navigation of the robot and the sound perception are disconnected, the SSL becomes difficult to perform with few microphones. Indeed, when considering an unknown motion of the sound source(s), erroneous localization are common because of the natural reverberation and the potential noise in the environment. The robustness of the system is gained from an array of microphones giving some redundant data [13]. In addition, complex post-processing methods have been developed to keep the track of the sound sources. For instance, a solution including eight microphones and a particle filter that tracks multiple sound sources is explored in [15]. Moreover, when the robot is moving in the scene, the tracking becomes even more difficult because of the additional dynamic noise that deteriorates the sound signal perception. In [15] the tracking is limited to two static sound sources when the robot is moving. Similarly the method proposed in [9] tracks one static sound source while using four microphones and a particle filter.

Yet controlling the robot accordingly to the perceived sound can improve the localization process as shown in [11]. In this latter work, an incremental control approach coupled with a neural network is used to track one static sound source with two microphones. Nonetheless the control input is determined beforehand by an operator. The idea of coupling motion and perception has also been exposed in binaural active audition [14]. This method takes into account the motion of the robot through an unscented Kalman filter tracking a sound source. However, this latter work has been limited so far to track only one sound source, and without any explicit model of the controller. In [7], the authors introduce a control scheme derived from a cost function characterizing the relationship between the position of a source and the stereo cues retrieved by a pair of microphones. Still this approach is specific to a particular task: namely the control of a rotational degree-of-freedom (DOF) to reach a unique orientation in an anechoic room.

In contrast, the sound-based control approach introduced in our previous work [8] processes the sound signal by straightly linking the aural perception of the environment, to the control of the robot in a sensor-based framework. The modelling is based on the measurement of the time difference of arrival (TDOA) between two microphones. The idea is to define a robotic task w.r.t angular information similarly to beaconing or bearing-only tasks. Thus our approach does not require an explicit localization of the sound sources. This paper consists in an experimental validation of the sound-based control theoretically presented in [8]. The experimental evaluation of our approach is performed considering the cases of one and two sound sources. We show that using such control scheme allows to decrease the number of microphones (2) w.r.t classical methods that usually require four to eight microphones. Indeed, by processing auditory

[1]Université Rennes I - IRISA, Campus de Beaulieu, 35042 Rennes cedex, France aly.magassouba@irisa.fr

[2]CNRS - IRISA, Campus de Beaulieu, 35042 Rennes cedex, France nancy.bertin@irisa.fr

[3]Inria - IRISA, Campus de Beaulieu, 35042 Rennes cedex, France francois.chaumette@inria.fr

cues directly in the control loop, our approach is more robust to reverberation and noise. Actually the sound-based control helps simplifying and improving the tracking of the sound sources. From the motion of the robot related to the TDOA, it is possible to predict the evolution of the TDOA measurements. Hence erroneous TDOA estimations due to reverberation or noise can be eliminated while a prediction model can replace missing measurements. Therefore the tracking method becomes simple compared to methods using particle filters or neural network.

The rest of this paper is organized as follows: we first introduce the sound-based control framework for one and two sound sources in Section II. A description of TDOA estimation and tracking is following in Section III. The paper ends in Section IV with experimental results validating this approach with one moving sound source and two static sound sources.

II. CONTROL MODELLING

A. Geometric configuration

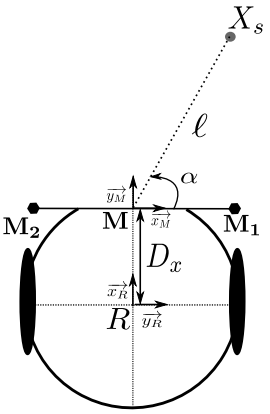


Fig. 1: Robot modelling

We consider a mobile robot modelled as a non-holonomic unicycle. Fig. 1 illustrates the design of the system instrumented by a pair of microphones located at $\mathbf{M_1}$ and $\mathbf{M_2}$. We set a frame $\mathcal{F}_{\mathbf{R}}(\overrightarrow{x_R}, \overrightarrow{y_R})$ attached to the center $\mathbf{R}$ of the robot. Besides a frame $\mathcal{F}_{\mathbf{M}}(\overrightarrow{x_M}, \overrightarrow{y_M})$ is defined so that its origin $\mathbf{M}$ is the midpoint of the pair of microphones. k sound source(s) $\mathbf{X_{s_i}}$ at a distance ℓ_i from $\mathbf{M}$ may be considered. In this paper we consider the case where $k = 1$ or $k = 2$. The sound emitted from each $\mathbf{X_{s_i}}$ reaches the microphones with an incident angle α_i in $\mathcal{F}_{\mathbf{M}}$. We also assume a far field condition so that the distance d between the microphones is small w.r.t each ℓ_i. D_x denotes the distance between the center of the robot $\mathbf{R}$ and $\mathbf{M}$. The robot is controlled upon two DOF: the control input $\dot{\mathbf{q}}$ is given by $(u\ \ \omega)$, respectively the translation velocity along $\overrightarrow{x_R}$ and the angular velocity around $\overrightarrow{z_R}$.

B. General framework

The task to realize consists in positioning the robot for satisfying given bearing conditions. This is performed by considering k TDOA measurements $\boldsymbol{\tau}(t)$ obtained from the microphones and by minimizing the error $\|\mathbf{e}(t)\|$ characterized by

$$\mathbf{e}(t) = \boldsymbol{\tau}(t) - \boldsymbol{\tau}^* \tag{1}$$

where $\boldsymbol{\tau}^*$ denotes the desired value of the TDOAs. The time variation of $\boldsymbol{\tau}$ is related to the sensors velocity by

$$\dot{\boldsymbol{\tau}} = \mathbf{L}_\tau \mathbf{v} \tag{2}$$

in which $\mathbf{L}_\tau \in \mathbb{R}^{k\times 3}$ is the interaction matrix sized by k the number of measurements and $\mathbf{v} = (v_x, v_y, \omega_z)$ denoting the spatial linear and angular velocity of the microphones expressed in $\mathcal{F}_{\mathbf{M}}$. This interaction matrix is obtained from the relationship between the TDOA and the sound source direction under the far field assumption given by

$$\tau = A\cos\alpha \tag{3}$$

where $A = d/c$ in which c is the sound celerity. The interaction matrix related to the TDOA has been determined in [8] as:

$$\mathbf{L}_\tau = \begin{bmatrix} -\frac{\nu^2}{A\ell} & \frac{\tau\nu}{A\ell} & \nu \end{bmatrix} \tag{4}$$

where $\nu = \sqrt{A^2 - \tau^2}$. Therefore, when considering the two-DOF robot previously described, the relationship between $\dot{\boldsymbol{\tau}}$ and the control input $\dot{\mathbf{q}}$ of the robot is:

$$\dot{\boldsymbol{\tau}} = \mathbf{J}_\tau \dot{\mathbf{q}} \tag{5}$$

where $\mathbf{J}_\tau$ corresponds more explicitly to:

$$\mathbf{J}_\tau = \mathbf{L}_\tau \mathbf{J}_\mathbf{r}. \tag{6}$$

$\mathbf{J}_r$ being the robot Jacobian given by

$$\mathbf{J_r} = \begin{bmatrix} 0 & D_x \\ 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{7}$$

Hence, it is possible design a control scheme given by:

$$\dot{\mathbf{q}} = -\lambda \widehat{\mathbf{J}_\tau^+} \mathbf{e} \tag{8}$$

where $\mathbf{J}_\tau^+ \in \mathbb{R}^{2\times k}$ is the Moore-Penrose pseudo-inverse of $\mathbf{J}_\tau$ and $\lambda > 0$ is a gain that tunes the time to convergence. Generally, an approximation $\widehat{\mathbf{J}_\tau^+}$ is considered since it is impossible to know perfectly either $\mathbf{J}_\tau$ or $\mathbf{J}_\tau^+$. Indeed the distance ℓ to the sound source used in (4) is a priori unknown. In practice we approximate the interaction matrix of τ with

$$\widehat{\mathbf{L}_\tau} = \begin{bmatrix} -\frac{\nu^2}{A\widehat{\ell}} & \frac{\tau\nu}{A\widehat{\ell}} & \nu \end{bmatrix}. \tag{9}$$

As shown in [3], this kind of approximation classical in Visual servoing does not degrade too much the system behaviour compared with using the real interaction matrix. The way to tune $\widehat{\ell}$ is discussed and given in the next section when considering one sound source and two sound sources.

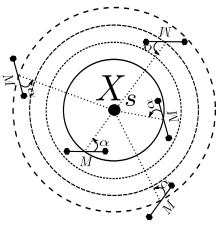


Fig. 2: With one sound source several poses centred on the sound source exists for a given τ^*

C. Sound-based control with one sound source

When considering more specifically one sound source in the environment, the interaction matrix of τ is given by (9). By using only one feature, it can be intuitively expected that several sensor poses exist so that $\tau = \tau^*$. Fig. 2 represents all possible poses that can fulfill the constraint $\tau = \tau^*$ (e.g $\alpha = \alpha^*$). These poses can be geometrically represented by a set of circles centred on the sound source $\mathbf{X_s}$. A task considering one sound source mainly specifies a desired orientation of the robot w.r.t the sound source. Therefore a control input characterized by the angular velocity ω only is sufficient to achieve any task involving one sound source. In that particular case where u would always be equal to 0 the Jacobian matrix reduces to

$$\widehat{\mathbf{J}_\tau} = \frac{A\widehat{\ell}\nu - D_x\nu^2}{A\widehat{\ell}} \tag{10}$$

and the control input becomes

$$\dot{q} = -\lambda \frac{1}{\nu\left(1 - \frac{D_x}{A\widehat{\ell}}\nu\right)}(\tau - \tau^*) \tag{11}$$

Excluding the degenerate case where $\alpha = 0$ or $\alpha = 180$ (i.e., $\nu = 0$), singularities of the control scheme could in principle occur if $\hat{\ell} < D_x$. Hopefully, this is impossible in practice. Indeed, knowing that $0 \leq \nu \leq A$, the denominator of (13) can never vanish as soon as $\hat{\ell} > D_x$. In the same way, it is very easy to demonstrate the global asymptotic stability of the system since the Lyapunov stability condition $\mathbf{J}_\tau\widehat{\mathbf{J}_\tau^+} > 0$ is satisfied as soon as $\ell > D_x$ and $\widehat{\ell} > D_x$.

This excellent stability property can also be obtained when the two DOF of the robot (u, ω) are controlled. The Jacobian matrix is then given by

$$\widehat{\mathbf{J}_\tau} = \begin{bmatrix} \frac{\tau\nu}{A\widehat{\ell}} & \frac{A\widehat{\ell}\nu - D_x\nu^2}{A\widehat{\ell}} \end{bmatrix} \tag{12}$$

and the control input is obtained as:

$$\dot{q} = -\lambda \begin{bmatrix} \frac{\tau\nu}{A}a_1 \\ -\nu(\frac{\nu D_x}{A} - \widehat{\ell})a_1 \end{bmatrix} (\tau - \tau^*) \tag{13}$$

where

$$a_1 = \frac{\tau^2 + (\nu D_x - A\widehat{\ell})^2}{A^2\widehat{\ell}^2}. \tag{14}$$

The control scheme input is not singular as soon as $\widehat{\ell} > 0$, which corresponds to the Lyapunov stability condition already demonstrated in [8]. So a relevant choice of $\widehat{\ell}$ consists in fixing this parameter to a positive and meaningful value.

D. Sound-based control with two sound sources

For two sound sources, the interaction matrix related to $\boldsymbol{\tau}$ is obtained by stacking (9) for each τ_i as:

$$\widehat{\mathbf{L}_\tau} = \begin{bmatrix} -\frac{\nu_1^2}{A\widehat{\ell_1}} & \frac{\tau_1\nu_1}{A\widehat{\ell_1}} & \nu_1 \\ -\frac{\nu_2^2}{A\widehat{\ell_2}} & \frac{\tau_2\nu_2}{A\widehat{\ell_2}} & \nu_2 \end{bmatrix}. \tag{15}$$

In this case, all the poses corresponding to a correct achievement of the task are such that $\mathbf{M}$ belongs to a circumscribed circle shaped by the two sound sources X_{s_1} and X_{s_2} with an orientation where $\tau_1 = \tau_1^*$ and $\tau_2 = \tau_2^*$ (see Fig. 3).

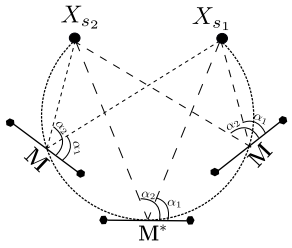


Fig. 3: With two sound sources several poses exist for given τ_1^* and τ_2^* on the circumscribed circle defined by the sound sources and the position of the microphones

For this task, the Jacobian of the system is given by:

$$\widehat{\mathbf{J}_\tau} = \begin{bmatrix} \frac{\tau_1\nu_1}{A\widehat{\ell_1}} & \frac{A\widehat{\ell_1}\nu_1 - D_x\nu_1^2}{A\widehat{\ell_1}} \\ \frac{\tau_2\nu_2}{A\widehat{\ell_2}} & \frac{A\widehat{\ell_2}\nu_2 - D_x\nu_2^2}{A\widehat{\ell_2}} \end{bmatrix} \tag{16}$$

and the control input by (8). A classical method to approximate each $\widehat{\ell_i}$ is to use the distance ℓ_i^* to the sound source at a desired pose. In this case, it is well known that the system is locally asymptotically stable in the neighborhood of the desired pose [3]. In our case, since the set of desired poses is infinite, many possible choices for $\widehat{\ell_i^*}$ are possible. We will see in Section IV how we have proceeded in practice.

Sound-based control can also be applied to three or more sound sources as theoretically exposed in [8]. In that case, only one single pose generally corresponds to the correct achievement of the task. Due to the non-holonomic constraints of a mobile robot, this would necessitate the design of a non-stationary control scheme, following the well-known Brockett theorem. This is of course outside the scope of this paper.

III. Sound processing

A. TDOA estimation

As exposed in the robot hearing literature [6], the estimation of the TDOA τ is performed by comparing two temporal signals $x_1(t)$ from the microphone $\mathbf{M_1}$ and $x_2(t)$ from $\mathbf{M_2}$ in the frequency domain with the Generalized cross-correlation (GCC):

$$\boldsymbol{R}_{1,2}(\tau) = \sum_f^F \frac{\phi_{x_1,x_2}(f)}{|\phi_{x_1,x_2}(f)|} e^{\varphi(\tau)}. \tag{17}$$

where φ corresponds to the phase shift for a defined τ and the PHAse Transform (PHAT) filter $\psi(f) = |\phi_{x_1,x_2}|^{-1}$ used in (17) is a normalization factor that increases the robustness towards reverberation. The cross-spectral power density ϕ_{x_1,x_2} is usually defined by

$$\phi^{sum}_{x_1,x_2}(f) = \frac{1}{L}\sum_{l=1}^{L} X_1(f,l)X_2^*(f,l) \tag{18}$$

for windows frames from l to L. $X_1(f,l)$ and $X_2^*(f,l)$ are respectively the Fourier transform of $x_1(t)$ and the conjugate of the Fourier transform of $x_2(t)$. In practice, we have preferred an alternative solution that consists in considering a "max" pooling function so that:

$$\phi^{max}_{x_1,x_2}(f) = \max_{l} \;\; X_1(f,l)X_2^*(f,l). \tag{19}$$

This solution gives the advantage of detecting sound sources active within few time frames by not integrating irrelevant information when these sources are inactive [2]. The maximum peak of the GCC function gives an estimation of the actual TDOA and can therefore be written as:

$$\widehat{\tau} = \underset{\tau}{\operatorname{argmax}} \; \boldsymbol{R}_{1,2}(\tau) \tag{20}$$

with $\tau \in [-A, A]$ corresponding to a sound direction of arrival α from 0° to 180°. When k sound sources are active, the GCC function returns several peaks wherein the k first peaks should correspond to the TDOA of each sound source. But, in real world conditions, the estimation of the TDOA(s) is altered by spurious peaks caused by reverberation and noise. Therefore p peaks ($p > k$) should be considered among which the k good peak(s) should be found.

B. Tracking routine

Similarly to other sensor-based approaches, the sound-based control is built upon a good tracking of the input features. Tracking the true value of the TDOA(s) among a set of observations is a major issue of the control scheme, since noise and reverberation affect the sound features measurement, with spurious and/or altered data. Yet, one of the benefit of coupling motion and perception lies in the prediction that limits the scope of erroneous measurements. Thus it is not necessary to use complex tracking methods to accurately perform a specified task. In our case, the tracking procedure is divided in two steps.

In the first step, the goal is to find the correct TDOA(s) in respect of the number of active sound sources defined beforehand. This step assumes that the robot and the sound source(s) are not moving. Then the set of peaks obtained from the GCC function is observed during a given number of windows frames. A clustering algorithm based on the Euclidean distance between the TDOAs is then applied to detect the frequency of appearance of a given TDOA τ_i. By combining this frequency to the mean appearance rank of each τ_i, the most probable TDOA(s) are then selected. This method is applied to the initial pose to retrieve $\tau_i(t_0)$ and could also be applied to retrieve each τ_i^* when the desired pose is not characterized by obvious TDOA(s). For some simple cases each τ_i^* can be defined manually, for instance for a task that consists in orienting the robot to the sound source (e.g., $\tau^* = 0$).

The second step is used during the motion of the robot, and consists in finding the genuine TDOA among the set of observations given by the GCC function. Knowing the previous value of the TDOA, we simply select the closest peak in the current frame. More evolved solutions could be implemented such as Kalman filter or Bayesian filter. They would provide more robust estimation of the correct TDOA but at a higher computation cost. Since the robustness of the sound-based control can cope with inaccurate and/or punctual errors in the estimation, we have preferred the simple and efficient solution exposed above for the experiments presented in Section IV.

Additionally to the tracking, a labelling of the retrieved TDOA is necessary when there are two sound sources. The goal is to associate each τ_i to the desired τ_i^* so that the task can be correctly completed. In the case of two sound sources, the labelling problem is trivial. Indeed, if we consider the working space as the half plane in front of the microphones, the ordinality of $\tau_i(t)$ and τ_i^* is the same. Namely if $\tau_1^* < \tau_2^*$ then $\tau_1(t)$ should be lesser than $\tau_2(t)$. As shown in Fig.3, it is obvious that each pose of the microphones in the environment can be characterized by the circumscribed circle on which the corresponding angles α_1 and α_2 have always the same order.

C. Overcoming missing measurements with a prediction model

By linking the features measurement to the control input, it is also possible to predict the evolution of the features in the next time frame. Given τ as the state x and the velocity $\dot{\mathbf{q}}$ applied to the robot, a local prediction model based on the Jacobian matrix (5) is simply given as follow:

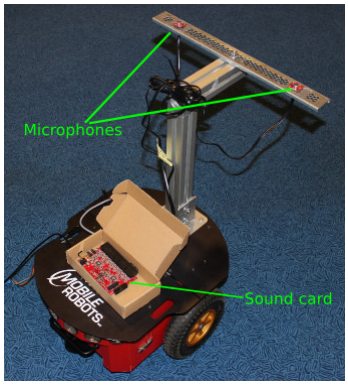
$$\begin{cases} \dot{x}(k) = \mathbf{J}_\tau \dot{\mathbf{q}} \\ x(k+1) = x(k) + T_e \dot{x}(k) \end{cases} \tag{21}$$

in which T_e refers to the sample time of the control loop. Nonetheless the predicted τ is not as accurate as the genuine TDOA, because of the approximation $\widehat{\mathbf{L}}_\tau$ used in (6) in which $\ell_i = \widehat{\ell}_i$. Several methods such as state space observer [4] could be used to obtain a better estimation of $\widehat{\ell}_i$, but the closed-loop control scheme is sufficiently robust to cope with a rough approximation of ℓ_i.

IV. EXPERIMENTAL RESULTS

A. Experimental setup

The experiments were conducted with a Pioneer 3DX robot on which two omnidirectional microphones are set as illustrated on Fig. 4. These microphones were connected to a sound card 8SoundsUSB [1] so that the sound was processed in real time. The sound card operates at a frequency of 48 kHz, and provides windows frames of 256 samples. The TDOA is computed from 10 consecutive windows frames (e.g 50 ms), that are sub-sampled at 16 kHz. Processing the sound signal at a frequency of 16 kHz gives two advantages:



d	0.31 m
c	343 m.s^{-1}
$\widehat{\ell}_i$	1 m
A	$0,00090379$ s
$\lambda(x)$	$5e^{(-4000x)}$

Fig. 4: Experimental settings

better results are obtained from uttered speeches by not considering high frequencies and the processing time is reduced with less samples to analyze. Consequently, the global control framerate is around 12 Hz. The environment consists in a room with a reverberation time (RT60) of approximately 580 ms. Moreover, the measured signal-to-noise ratio (SNR) is around 20 dB in presence of typical noise such as computer noise and ventilation in the room. We processed two experiments detailed in the next sections and in Fig. 5. The parameters given in Fig. 4 are used for both experiments. An adaptive gain $\lambda(x)$ in which x refers to the infinity norm of the error $\mathbf{e}$ is used to smooth the robot motion. The accompanying video to this paper illustrates the results obtained.

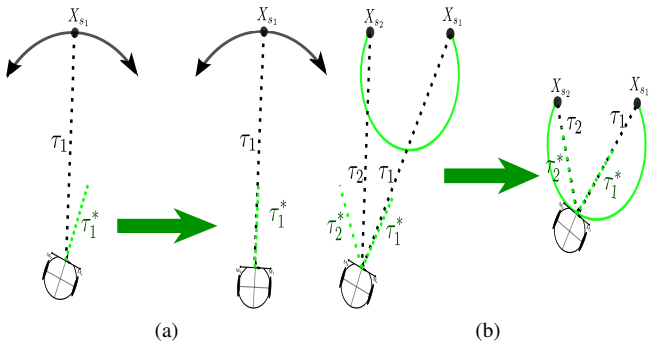
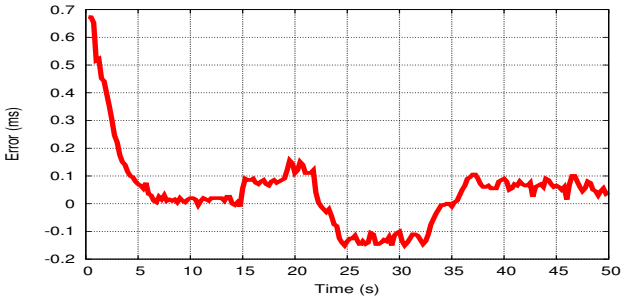


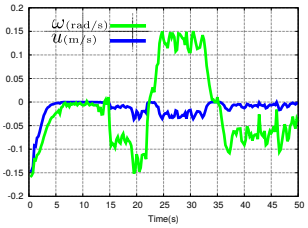
Fig. 5: In experiment (a), one moving sound source is considered, the goal is to maintain the microphones in the direction of the sound source ($\tau^* = 0$). In experiment (b), two still sound sources are considered and the goal is to reach the circle that defines the set of solution where $\tau_1 = \tau_1^*$ and $\tau_2 = \tau_2^*$.

B. Sound-based control with one moving sound source

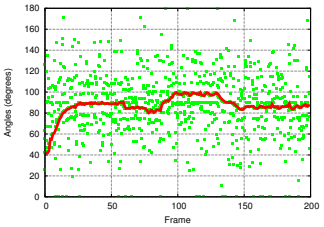
In this experiment we considered one sound source that corresponds to a female voice recording of 10s played in loop. The goal of the experiment is to maintain the robot oriented in the direction of the sound source. In the first step the robot is randomly oriented and τ^* is set to 0, while the sound source is 1.5 m away from the robot. In this part of the experiment the robot correctly positioned itself in the direction of the static sound source, as illustrated by the exponential decrease of the error during the 5 first seconds in Fig.6a. Subsequently the sound source was moved from one side in the environment and at different distances. As a result, the robot constantly moved in the direction of the source, despite spurious measurements due to reverberation or noise. Indeed we can notice in Fig. 6c several false measurements represented by the green dots not tracked by the system. The effect of the noise can also be denoted by the aligned green dots at $\alpha = 90°$ during all the frames of the task. Nonetheless, the task is correctly achieved since the lag between the robot orientation and the desired one did not exceed $10°$ despite the low dynamic response of the mobile robot. Nevertheless this tracking error could be reduced by introducing an integrator in the control law or more advanced control techniques [3].



(a) Error $\tau - \tau^*$



(b) Control input



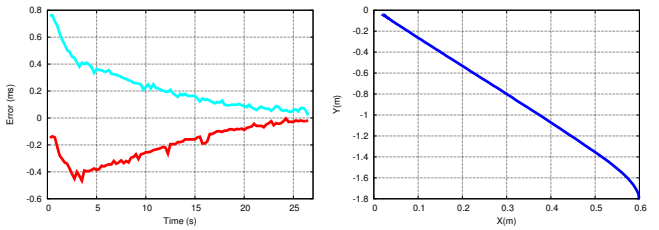
(c) α observation from TDOA (dots) and tracking (line)

Fig. 6: Experiment with the robot tracking one moving sound source.

C. Sound-based control two static sound sources

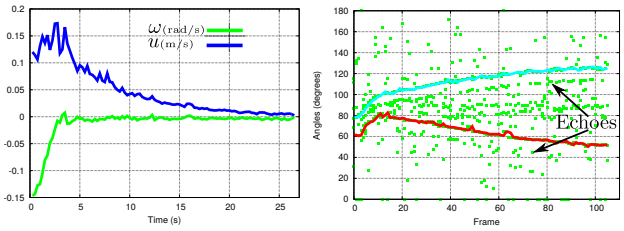
In this experiment, besides the female speech we added a second sound source corresponding to a burst of white Gaussian noise of 25 ms followed by 25 ms of silence played in loop. This time, the objective is to reach a pose where $\tau_1^* = -\tau_2^*$ with $\alpha_1 = 50°$. From a pose fulfilling that condition, the system extracted τ_1^* and τ_2^* in the first step. After, starting from a pose around 3 meters away from the sources, the system automatically initialized and labelled $\tau_1(t_0)$ and $\tau_2(t_0)$. The results in Fig. 7 show a correct initialization followed by the completion of the task. Once again, corrupted TDOA estimations occurred during the robot motion but were coped with the tracking routine. More precisely the spurious TDOAs were caused by echoes following the same dynamic as the real TDOAs. This could

be expected since the echoes appear as virtual sound sources. Moreover the noise effect is still present with observation of peaks for $\alpha = 90°$ during most of the frames. Despite these poor observations, the error of the measured TDOAs successfully converges to zero while the robot followed a straight and smooth trajectory in the direction of the circle to be reached and with a correct orientation.



(a) Error $\boldsymbol{\tau} - \boldsymbol{\tau}^*$

(b) Trajectory from the robot odometry



(c) Control input

(d) α observation from TDOA (dots) and tracking (lines)

Fig. 7: Experiment with two sound sources

V. CONCLUSION

In this paper, the sound-based control approach proposed in our previous work [8] has been validated for the first time with experiments on a mobile robot. The obtained results show the effectiveness and the benefits of this method in the case of one moving and two fixed sound sources. First, this method is robust to reverberation and noise since this type of sensor-based approach can cope with false and missing measurements. Punctual errors and approximation in the TDOA estimation do not compromise the good progress of the task. Therefore this approach reduces the complexity of TDOA pre-processing computation (for instance noise filtering or dereverberation) besides simplifying the TDOA tracking process.

In the case of one sound source, we have been able to correctly orient the robot in the direction of one moving sound source. This task was performed in real time. This result is a clear contribution to robot hearing since we obtained similar results to array-based localization without complex filtering methods such as Kalman or particle filter while using only two microphones. Therefore applying this method to binaural hearing which is known as less robust than array-based localization can overcome this flaw.

In the second experiment, we have been able to correctly position the robot with respect to two sound sources. The robot successfully reached a pose characterized by the desired bearing conditions. Furthermore the tracking of the two sound sources were correctly performed among a set of spurious and altered TDOA estimations.

This kind of positioning task can have many uses for multi-robot applications. Indeed when considering robots that have to move according to the rest of the group, taking into account the hearing sense seems particularly suitable.

Ongoing work concerns the use of different auditory cues in a similar sensor-based control framework. Interaural phase difference, interaural level difference or sound energy are also features to prospect in order to achieve different and various tasks. Experimental extension to more sound sources and different robots is also intended in a near future.

REFERENCES

[1] D. Abran-Côté, M. Bandou, A. Béland, G. Cayer, S. Choquette, F. Gosselin, F. Robitaille, D. T Kizito, F. Grondin, and D. Létourneau. Usb synchronous multichannel audio acquisition system. Technical report, 2014.

[2] C. Blandin, A. Ozerov, and E. Vincent. Multi-source tdoa estimation in reverberant audio using angular spectra and clustering. *Signal Processing*, 92(8):1950–1960, 2012.

[3] F. Chaumette and S. Hutchinson. Visual servoing and visual tracking. In *Springer Handbook of Robotics*, pages 563–583. Springer, 2008.

[4] A. De Luca, G. Oriolo, and P. Robuffo Giordano. Feature depth observation for image-based visual servoing: Theory and experiments. *The Int. Journal of Robotics Research*, 27(10):1093–1116, 2008.

[5] A. Deleforge, F. Forbes, and R. Horaud. Variational em for binaural sound-source separation and localization. In *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 76–80, 2013.

[6] C.H Knapp and G.C Carter. The generalized correlation method for estimation of time delay. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 24(4):320–327, 1976.

[7] M. Kumon, T. Sugawara, K. Miike, I. Mizumoto, and Z. Iwai. Adaptive audio servo for multirate robot systems. In *IEEE Int. Conf. on Intelligent Robots and Systems*, volume 1, pages 182–187, 2003.

[8] A. Magassouba, N. Bertin, and F. Chaumette. Sound-based control with two microphones. In *IEEE Int. Conf. on Intelligent Robots and Systems*, 2015.

[9] I. Marković and I. Petrović. Speaker localization and tracking with a microphone array on a mobile robot using von mises distribution and particle filtering. *Robotics and Autonomous Systems*, 58(11):1185–1196, 2010.

[10] J C. Murray, H. Erwin, and S. Wermter. Robotics sound-source localization and tracking using interaural time difference and cross-correlation. In *Proc. of NeuroBotics Workshop*, pages 89–97, 2004.

[11] J C. Murray, H. Erwin, and S. Wermter. Robotic sound-source localisation architecture using cross-correlation and recurrent neural networks. *Neural Networks*, 22(2):173–189, 2009.

[12] T. Nakadai, K.and Lourens, H G. Okuno, and H. Kitano. Active audition for humanoid. In *AAAI/IAAI*, pages 832–839, 2000.

[13] H. Nakajima, K. Kikuchi, T. Daigo, Y. Kaneda, K. Nakadai, and Y. Hasegawa. Real-time sound source orientation estimation using a 96 channel microphone array. In *IEEE Int. Conf. on Intelligent Robots and Systems*, pages 676–683, 2009.

[14] A. Portello, P. Danes, and S. Argentieri. Active binaural localization of intermittent moving sources in the presence of false measurements. In *IEEE Int. Conf. on Intelligent Robots and Systems*, pages 3294–3299, 2012.

[15] J-M. Valin, F. Michaud, and J. Rouat. Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering. *Robotics and Autonomous Systems*, 55(3):216–228, 2007.

[16] Q H. Wang, T. Ivanov, and P. Aarabi. Acoustic robot navigation using distributed microphone arrays. *Information Fusion*, 5(2):131–140, 2004.

[17] J. Woodruff and D. Wang. Binaural localization of multiple sources in reverberant and noisy environments. *IEEE Trans. on Audio, Speech, and Language Processing*, 20(5):1503–1512, 2012.